Alternative posterior consistency results in nonparametric binary regression using Gaussian process priors

Taeryon Choi

Department of Mathematics and Statistics,
University of Maryland, Baltimore County

BNRW01 Aug. 09, 2007

Outline

Gaussian Process Priors

- Bayesian nonparametric estimation of $f(x)$
  - Bayesian analysis with infinite-dimensional parameter space
  - Prior probability distribution for random function $f(x)$
  - $f(x) = g(\eta(x))$, where $\eta(\cdot) \sim GP(\mu(\cdot), R(\cdot,\cdot)$ and $g(\cdot)$ is assumed to be known.
- A stochastic process is a random function, regarded as a single random variable taking values in an infinite-dimensional function space
- Provides a natural way of defining prior distributions over spaces of functions
- S. Petrone, CKI Williams, A. Simoni, D. Cox, J.Q.Shi ...

Posterior Consistency in Regression Problems and Gaussian Process Priors

- Choi (2007) : binary regression with Gaussian process priors
- Tokdar and Ghosh (2007) : density estimation with Gaussian process priors
- Choi and Schervish (2007) : nonparametric regression with Gaussian errors
- Ghosal and Roy (2006) : binary regression with Gaussian process priors
- Coram and Lally (2006) : binary regression (uniform mixture prior)
- Choi (2005) : nonparametric regression under Gaussian proces priors
- Walker (2003) : Nonparametric regression (both binary regression and nonparametric regression with Gaussian errors)

Posterior Consistency

The sequence of posterior distribtuions $\{\Pi(\cdot|X_1, X_2, \ldots, X_n)\}$ is said to be consistent at θ_0, if the posterior for P_{θ_0} almost all sequences of observations converges, in a suitable sense, to the degenerate measure at θ_0.

- A kind of frequentist validation of the updating method.
- If an oracle were to know the true value of the parameter, posterior consistency ensures that with enough observations one would get close to this true value.
- There are other intepretations related to merging of opinions and other concepts.

Posterior Consistency with L_1 neighborhood

- The sequence $\{\Pi_n(\cdot|X^{(n)})\}$ is said to be consistent at θ_0 if for every neighborhood U of θ_0,

$$\Pi_n(U^C|X^{(n)}) \stackrel{n\to\infty}{\longrightarrow} 0 \;\; a.s.\ [P_{\theta_0}]$$

- Consider the L_1 neighborhood $U = \{\theta \;:\; \|f_\theta - f_{\theta_0}\| < \epsilon\}$
- Sufficient conditions
  - Ghosal, Ghosh and Ramamoorthi (1999) (GGR)
  - Barron, Schervish and Wasserman (1999) (BSR)
  - Walker (2004) (W)

Sufficient Conditions for Posterior Consistency

- θ_0 is in the KL support of Π
- In addition,

(W) For each $\delta > 0$, there exist sets $A_1, A_2, \ldots$ such that $\cup A_i = \Theta$, L_1-diameter of $\{f_\theta : \theta \in A_i\} < \delta$ and $\sum_i \sqrt{\Pi(A_i)} < \infty$

(BSW) For each $\epsilon > 0$, there exist $\Theta_n \subset \Theta$, and C, c_1, c_2, δ all positive such that

1. $\Pi(\Theta_n^c) < e^{-nc_2}$
2. $\mathcal{H}(\Theta_n, \delta) \le nc$ for $c < ([\epsilon - \sqrt{\delta}]^2 - \delta)/2, \delta < \epsilon^2/4$

(GGR) If for each $\epsilon > 0$, there is a $0 < \delta < \epsilon, c_1, c_2, \beta < \epsilon^2/2$ and Θ_n such that

1. $\Pi(\Theta_n^c) < c_1 e^{-n\beta}$
2. $J(\Theta_n, \delta) \le n\beta$

- $(W) \Rightarrow (GGR)$ and $(BSW) \Rightarrow (GGR)$

- $(W) \Rightarrow (GGR)$:

$$\Theta_n = \bigcup_{1}^{k_n} A_i.$$

1. $J(\Theta_n, \delta) < \log k_n$, taking $k_n = e^{n\beta}$
2. $\Pi(\Theta_n^c) = \Pi\left(\bigcup_{i>k_n} A_i\right) \leq \frac{2c^2}{k_n}$

- $(BSW) \Rightarrow (GGR)$: $J(\Gamma, 2\delta) \leq \mathcal{H}(\Gamma, \delta)$

- Choi and Ramamoorthi (2007)

Extension of Posterior consistency for Non i.i.d

1. A stronger condition related to KL support, for prior positivity
2. Existence of uniformly consistent tests
   - Ameou-Atisso, Ghosal, Ghosh and Ramamoorthi (2003)
   - Choudhuri, Ghosal and Roy (2004)
   - Choi and Schervish (2007)
   - Choi (2007)
3. Walker's sufficient conditions are easily adaptable
   - Walker (2003)
   - Choi and Ramamoorthi (2007)

1. For a probability measure ν on θ, let $q_\nu^{(n)}$ be the marginal density of $X_1, \ldots, X_n$, $q_\nu^{(n)}(x_1, x_2, \ldots, x_n) = \int_\Theta f_\theta^{(n)}(x_1, x_2, \ldots, x_n)\nu(d\theta)$.
2. Let $A \subset \Theta$ and $\delta > 0$. A and θ_0 are said to be *strongly δ separated* if for any probability ν on A, $\mathrm{Aff}(f_{\theta_0}, q_\nu^{(1)}) < \delta$
3. $\mathrm{Aff}(f, g) = \int \sqrt{fg}\, d\mu$, $H^2(f, g) = 1 - 2\mathrm{Aff}(f, g)$

If $A = \bigcup_{i\geq 1} A_i$ such that

1. For some $\delta > 0$ all the A_i's are strongly δ separated from θ_0 for the model $\theta \mapsto f_{i,\theta}$ and
2. $\sum_{i\geq 1} \sqrt{\Pi(A_i)} < \infty$

Then for some $\beta_0 > 0$,

$$e^{n\beta_0} \int_A \prod_{i=1}^{n} \frac{f_{i,\theta}(x_i)}{f_{i,\theta_0}(x_i)} \Pi(d\theta) \to 0, \text{ a.s } \prod_{i=1}^{\infty} P_{i,\theta_0}.$$

Theorem 2 of Choudhuri, Ghosal and Roy (2004)

(A1) Prior positivity of neighborhoods. Suppose that there exists a set B with $\Pi(B) > 0$ such that

(i) $\displaystyle \frac{1}{r_n^2} \sum_{i=1}^{r_n} V_{i,n}(\theta_0, \theta) \to 0$ for all $\theta \in B$,

(ii) $\displaystyle \lim_{n\to\infty} \inf \Pi \left(\left\{ \theta \in B : \frac{1}{r_n} \sum_{i=1}^{r_n} K_{i,n}(\theta_0, \theta) < \epsilon \right\} \right) > 0$ for all $\epsilon > 0$,

(A2) Existence of tests
Suppose that there exists test functions $\{\Phi_n\}$, subsets $\Theta_n \subset \bar{\Theta}_n$ and constants $C_1, C_2, c_1, c_2 > 0$ such that

(i) $\mathrm{E}_{\theta_0}\Phi_n \to 0$

(ii) $\sup_{\theta \in U_n^c \cap \Theta_n} \mathrm{E}_\theta(1 - \Phi_n) \le C_1 e^{-c_1 r_n}$,

(iii) $\Pi(\bar{\Theta}_n \cap \Theta_n^c) \le C_2 e^{-c_2 r_n}$.

Basic Setup

$$
\begin{aligned}
Y_i|x_i &\sim \mathrm{Binomial}(1, p(x_i)), \quad i = 1, \dots, n, \\
p(x) &= H(\eta(x)), \\
\eta(x) &\sim GP(\mu(\cdot), R(\cdot, \cdot)).
\end{aligned}
$$

- x_i's are fixed in advance or sampled from a probability distribution Q on the compact set $T \in \mathbb{R}^d$.
- $\eta(x)$ is assumed to be a Gaussian process parameterized by its mean function $\mu : T \rightarrow \mathbb{R}$ and its covariance function $R : T^2 \rightarrow \mathbb{R}$, denoted by $GP(\mu, R)$.
- The true probability function $p_0(x) = H(\eta_0(x))$ is supposed to be a function of the covariate x,
- $\eta_0(x)$ has a continuously differentiable sample path on T.
- $T = [0, 1]^d$

Assumptions of Gaussian Process Priors

P1. For all $n \geq 1$, all $\beta > 0$ and all $x_1, \ldots, x_n \in [0,1]$, the n-variate covariance matrix, $((\Sigma_{i,j}))$ with $\Sigma_{i,j} = R(x_i, x_j; \beta)$, is non-singular.

P2. The covariance function, $R(x, x'; \beta)$ has the form $R_0(\beta|x - x'|)$, where $R_0(x)$ is a positive multiple of a nowhere zero density function on $\mathbb{R}$ and four times continuously differentiable on $\mathbb{R}$.

P3. The mean function $\mu(x)$ of the Gaussian process $\eta(x)$ is continuously differentiable in $[0,1]$.

P4. There exists $0 < \delta < 1/2$ and $b_1,\ b_2 > 0$ such that

$$\kappa\left\{\beta > n^{\delta}\right\} = \Pr\left\{\beta > n^{\delta}\right\} < b_1 \exp(-b_2 n), \quad \forall n \geq 1$$

Assumptions of Gaussian Process Priors : d-dimensional

P2$_d$ Let $\tilde{\beta} = (\beta_1, \ldots, \beta_d)$. The covariance function, $R(\mathbf{x}, \mathbf{x}'; \tilde{\beta})$ is a product of d isotropic and integrable covariance functions, one for each dimension.

$$R(\mathbf{x}, \mathbf{x}'; \tilde{\beta}) = R^{(1)}(x_1, x_1'; \beta_j) R^{(2)}(x_2, x_2'; \beta_2) \ldots R^{(d)}(x_d, x_d'; \beta_d),$$

where each $R^{(i)}(x_i, x_i'; \beta_i) = R_{0,i}(\beta_i |x_i - x_i'|)$, where $R_{0,i}$ is a positive multiple of density.

P3$_d$ The mean function $\mu(\mathbf{x})$ of the Gaussian process $\eta(\mathbf{x})$ is continuously differentiable and $R_0(x)$ has continuous partial derivatives up to order $2d + 2$.

P4$_d$ β_j has a prior distribution, κ_j, with support $\mathbb{R}^+$, and there exists $0 < \delta < 1/2$ and $b_1,\ b_2 > 0$ such that such that

$$\kappa \left\{ \beta_j > n^{\delta} \right\} = \Pr \left\{ \beta_j > n^{\delta} \right\} < b_1 \exp(-b_2 n),\ \ \forall n \geq 1,\ j = 1, \ldots, d.$$

Alternative posterior consistency results

- Ghosal and Roy (2006) : the existence of the first and the second continuous sample path derivatives.
- Assumption P only ensures the existence of continuous sample path derivative of Gaussian process,
- $$\Theta_n = \{p(\cdot) : p(x) = H(\eta(x)),\ \|D^w\eta\|_\infty < M_n, |w| \leq 1\}, \qquad (1)$$

  where $D^w\eta = (\partial^{|w|}/\partial^{w_1}x_1 \ldots \partial^{w_d}x_d)\eta(x_1, \ldots, x_d)$, $|w| = \sum w_j$ and $M_n = O(n^{\alpha_1})$ and $\dfrac{2\delta+1}{2} < \alpha_1 < 1$ for some $0 < \delta < 1/2$.
- the transformed true response function $\eta_0(x)$ is assumed to be continuously differentiable.

Alternative Consistency Results (Cont'd)

- Based on the usual L_1 metric between two probability functions.
- An intermediate metric for two probability functions, the "in-measure metric"

$$d_Q(f,g) = \inf\{\epsilon : Q(\{x : |f(x) - g(x)| > \epsilon\}) < \epsilon\}. \tag{2}$$

- Fixed covariates :
  $\Pi\left\{\int |p(x) - p_0(x)|dx > \epsilon \middle| Y_1, \ldots, Y_n, x_1, \ldots x_n\right\} \to 0$ in P_0^n-probability.
- Random covariate :
  $\Pi\left\{\int |p(x) - p_0(x)|dQ(x) > \epsilon \middle| (X_1, Y_1) \ldots (X_n, Y_n)\right\} \to 0$ in P_0^n-probability.

Verifications

- Prior positivity conditions
  - $\Pi(\eta : \|\eta - \eta_0\|_\infty < \epsilon) > 0$ for every $\epsilon > 0$ when the link function H is assumed to be bounded and Lipschitz continuous.
  - the uniform support of a Gaussian process which has been thoroughly examined by Tokdar and Ghosh (2007) and Ghosal and Roy (2006)
  - the prior positivity condition holds under Assumption P,
- Existence of Tests
  - We consider a similar test to that of Ghosal and Roy (2006) but with a different technique and a weaker condition.
  - Choi and Schervish (2007)
  - Type I and II errors are exponentially small

- Let p_1 be a continuous function on $[0,1]$ and define $p_{ij} = p_i(x_j)$ for $i = 0, 1$ and $j = 1, \ldots, n$. Let $\epsilon > 0$, and let $r > 0$. Let $c_n = n^{\tau_1}$ for $\alpha_1/2 < \tau_1 < 1/2$ and $1/2 < \alpha_1 < 1$. Let $b_j = 1$ if $p_{1j} \geq p_{0j}$ and -1 otherwise. Let $\Psi_n[p_1, \epsilon]$ be the indicator of the set A_1, where A_1 is defined as

$$A_1 = \left\{ \sum_{j=1}^{n} b_j \left(Y_j - p_{0j} \right) > 2 c_n \sqrt{n} \right\}, \text{ where } Y_j \sim Bernoulli(p_{0j}).$$

- Then there exists a constant C_3 such that for all p_1 that satisfy

$$\sum_{j=1}^{n} |p_{1j} - p_{0j}| > rn, \tag{3}$$

$\mathrm{E}_{P_0}(\Psi_n[\eta_1, \epsilon]) < C_3 \exp(-2c_n^2)$. Also, there exist constants C_4 and C_5 such that for all sufficiently large n and all p satisfying $\|p - p_1\|_\infty < r/4$,

$$\mathrm{E}_P(1 - \Psi_n[p_1, \epsilon]) \leq C_4 \exp(-nC_5),$$

where P is the joint distribution of $\{Y_n\}_{n=1}^{\infty}$ assuming that $\theta = p$.

- Bernstein's inequality

- Fixed covariates :

  Let Q be the Lebesgue measure. Let $V > 0$ be a constant. For each integer n, let A_n be the set of all continuous functions γ such that $\forall x_1,\ x_2 \in [0,1]$, $|\gamma(x_1) - \gamma(x_2)| \le (M_n + V)|x_1 - x_2|$, where M_n is defined in (1). For each function γ and $\epsilon > 0$, define $B_{\epsilon,\gamma} = \{x : |\gamma(x)| > \epsilon\}$. Then for each $\epsilon > 0$ there exists an integer N such that, for all $n \ge N$ and all $\gamma \in A_n$,

$$\sum_{i=1}^{n} |\gamma(x_i)| \ge \ nQ(B_{\epsilon,\gamma})\frac{\epsilon}{3}. \tag{4}$$

- Random covariate :

  Let p be a function such that $d_Q(p, p_0) > \epsilon$. Let $0 < r < 2$ be a constant, and define

$$A_n = \left\{ \sum_{i=1}^{n} |p(X_i) - p_0(X_i)| \ge rn \right\}.$$

  Then there exists $C_1 > 0$ such that $\Pr(A_n^C) \le \exp(-C_1 n)$ for all n and A_n occurs all but finitely often with probability 1. The same C_1 works for all p such that $d_Q(p, p_0) > \epsilon$.

Examples of Covariance Functions

- $R_0(x)$ is a positive multiple of nowhere zero density function
- $R_0(x)$ is four times continuously differentiable
- Examples of $R_0(x)$
  - Squared-exponential : $R_0(x) = \exp(-x^2)$
  - Cauchy : $R_0(x) = \frac{1}{1+x^2}$
  - Matérn covariance function with $v > 2$ $R_0(x) = \dfrac{1}{2^{v-1}}(\alpha x)^v K_v(\alpha x)$, where $\alpha > 0$ and $K_v(x)$ is a modified Bessel function of order v.

Hyperparameters of Covariance Function

- An example of a covariance function, $R_\lambda(s,t;\beta)$
  - The squared exponential covariance function :
    $R_\lambda(s,t;\beta) = \lambda \exp\left(-\frac{\beta^2(s-t)^2}{2}\right)$
- We also establish posterior consistency with additional hyperparameters.
  - $\Pi\left\{ U_n^C \middle| Y_1,\ldots,Y_n,x_1,\ldots,x_n \right\} \to 0,\ a.s.\ [P_{\theta_0}]$
  - Compactness of A_Λ
  - Continuity as a function of λ in the covariance function.

Summary

- A theoretical justification of GP binary regression in terms of posterior consistency
  - Extension of the results of Ghosal and Roy (2006)
  - Under a weaker smoothness condition
- A comparison of various sufficient conditions for posterior consistency
- Extension of posterior consistency to non i.i.d setting

Future Work

- Other immediate(?) asymptotic issues :
  - Applying Walker's sufficient conditions to regression problems
  - Relationships among sufficient conditions for the convergence rate : Shen and Wasserman (2001), Ghosal, Ghosh and van der Vaart (2000), Walker, Lijoi and Prünster (2007)
  - Extension to non iid setting : Ghosal and van der Vaart (2007)
  - Existence of uniformly exponentially consistent tests under non iid setup : Barron (1989)

- Gaussian Process priors :
  - Rate of convergence of posterior distribution : van der Vaart and van Zanten (2007)
  - Small ball probability of Gaussian processes

References

1. Choi, T. (2007). Alternative posterior consistency results in nonparametric binary regression using Gaussian process priors. J. Statist. Plann. Inference. 137, 2975–2983
2. Choi, T. and Ramamoorthi, R.V. (2007). Remarks on consistency of posterior distributions, unpublished mansuscript
3. Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. J. Multivariate Anal. *to appear*
4. Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. Ann. Statist. 34, 2413–2429
5. Ghosal, S. and Van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. Ann. Statist. 35, 192–223
6. Tokdar, S. and Ghosh, J. K. (2007). Posterior consistency of Gaussian process priors in density estimation. J. Statist. Plann. Inference. 137, 34–42
7. Van der Vaart, A. W. and Van Zanten, J. H. (2007). Rates of contration of posterior distributions based on Gaussian process priors. Ann. Statist. *to appear*
8. Walker, S. G. (2004). New approaches to Bayesian consistency. Ann. Statist. 32, 2028–2043
9. Walker, S. G., Lijoi, A., and Prüster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. Ann. Statist. 35, 738–746